University of Virginia School of Law

Public Law and Legal Theory Research Paper Series 2016-10
February 2016

Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men

by

Jennifer L. Skeem
University of California, Berkeley
John Monahan
University of Virginia School of Law
Christopher T. Lowenkamp
Administrative Office, U.S. Courts

DRAFT: January 18, 2016

Gender, risk assessment, and sanctioning:

The cost of treating women like men

Jennifer Skeem

University of California, Berkeley

John Monahan

University of Virginia School of Law

Christopher Lowenkamp

Administrative Office, U.S. Courts

Wordcount:

Corresponding author: Jennifer Skeem, University of California, Berkeley, 120 Haviland Hall #7400, Berkeley, CA 94720-7400

Note: The views expressed in this article are those of the authors alone and do not reflect the official position of the Administrative Office of the U.S. Courts.

Abstract

Increasingly, jurisdictions across the U.S. are using risk assessment instruments to scaffold efforts to unwind mass incarceration without compromising public safety. Despite promising results, critics oppose the use of these instruments to inform sentencing and correctional decisions. One argument is that the use of instruments that include gender as a risk factor will discriminate against men in sanctioning. Based on a sample of 14,310 federal offenders, we empirically test the predictive fairness of an instrument that omits gender, the Post Conviction Risk Assessment (PCRA). We found that the PCRA strongly predicts arrests for both genders—but overestimates women's likelihood of recidivism. For a given PCRA score, the predicted probability of arrest—which is based on combining both genders—is too high for women. Although gender neutrality is an obviously appealing concept, it may translate into instrument bias and overly harsh sanctions for women. With respect to the moral question of disparate impact, we found that women obtain slightly lower mean scores on the PCRA than men (*d*= .32); this difference is wholly attributable to men's greater criminal history, a factor already embedded in sentencing guidelines.

Key words: risk assessment, gender, test bias, disparities, sentencing

Gender, risk assessment, and sanctioning:

The cost of treating women like men

Crime and criminal justice in early 21st century America are characterized by two events. One event is the rise of “mass incarceration.” The growth in incarceration rates in the United States since the early 1970s has been “historically unprecedented and internationally unique” (Travis, Western, & Redburn, 2014, p. 2). Approximately one percent of the adult American population—2.3 million people—now reside in jails or prisons (Sabol, West, & Cooper, 2009). Western European democracies have an incarceration rate one-seventh that of the United States (International Centre for Prison Studies, 2013). The costs associated with what some have called the “carceral state” (Simon, 2007) have become morally indefensible and fiscally unsustainable (Cullen, Jonson, & Nagin, 2011).

The other event that characterizes modern America is that the crime rate has plummeted in recent decades. The number of violent crimes committed per 100,000 Americans was 758 in 1991 and 365 in 2014—a decrease of 52% (FBI, 2015). In some cities, the crime decline has been nothing short of astounding. In New York City, for example, the homicide rate now is 18% of what it was in 1990 (Zimring, 2012). Some have suggested that these two events are strongly causally related—i.e., that the rise in the rate of imprisonment largely produced the fall in the rate of crime (Rushford, 2015). However, an extensive examination of this hypothesis by the National Research Council concluded that “the increase in incarceration may have caused a decrease in crime, but the magnitude of the reduction is highly uncertain and the results of most studies suggest it was unlikely to have been large” (Travis, Western, & Redburn, 2014, p. 4).

Advocates across the political spectrum (Arnold & Arnold, 2015) have seized on the juxtaposition of these two events to argue that it may be possible to reduce mass incarceration without simultaneously increasing crime. A principal strategy proposed to accomplish this goal is the incorporation of risk assessment throughout the sanctioning process. More specifically, advocates envision three ways that risk assessment can reduce jail and prison populations without increasing the crime rate (Monahan & Skeem, in press). First, risk assessment can inform decisions about whether an offender has *a sufficiently high likelihood of again committing crime* to justify a period of incapacitation. That is, within a range of severity set by moral concerns about the blameworthiness of the offender for committing the crime of which he or she has been convicted, risk assessment can inform decisions about whether—on utilitarian crime-control grounds—an offender should be sentenced to the upper-bound of that range (Skeem & Monahan, 2011).

Second, risk assessment can inform decisions about whether an offender has *a sufficiently low likelihood of again committing crime* to justify an abbreviated period of incapacitation or, in the case of supervised probation or parole, no incapacitation at all. That is, within a range of severity set by moral concerns about the blameworthiness of the offender for committing the crime of which he or she has been convicted, risk assessment can inform decisions about whether—on utilitarian crime-control grounds—an offender should be sentenced to the lower-bound of that range (Monahan & Skeem, 2014).

Finally, risk assessment can inform the type and intensity of correctional interventions designed *to reduce an offender's likelihood of again committing crime.* That is, risk assessment instruments can be used to identify "causal" risk factors that can be modified by evidence-based interventions. To the extent that such causal risk factors for crime can be identified and modified, risk assessment can do more than passively estimate an offender's likelihood of recidivism. It can actively reduce that likelihood (Lowenkamp, Latessa, & Holsinger, 2006; Dvoskin, Skeem, Novaco, & Douglas, 2012). These correctional interventions might take place while an offender is incapacitated in a jail or a prison, or is under legal restraint in the community on probation or parole.

While many extoll the potential contribution of risk assessment to reducing mass incarceration without increasing crime, others are equally adamant in opposition to incorporating risk assessment in the sanctioning process. The principal concern is that any benefits in terms of reduced rates of incarceration achieved through the use of risk assessment will be offset by costs to social justice claimed to be inherent in the risk assessment enterprise. Based on a sample of over 34,000 offenders, Skeem and Lowenkamp (2015) have examined this claim with respect to race and one well-known risk assessment instrument—the federal Post Conviction Risk Assessment [PCRA] (Johnson, Lowencamp, VanBenschoten & Robinson, 2011). There were two relevant findings. First, the PCRA was free of *predictive bias*—the instrument predicted re-arrest for both African American and white offenders strongly, and with similar form (i.e., a given PCRA score roughly corresponded to a given probability of recidivism, across race). Second, on average, African American offenders obtained modestly higher PCRA scores than white offenders (mostly because of higher scores on the criminal history scale). Although these differences do not reflect test bias, some uses of the PCRA could have *disparate impact* on African American offenders.

In this study, we use a comparable sample of federal offenders to examine questions regarding the same instrument (the PCRA) as Skeem and Lowenkamp, but with a focus on the role played by gender rather than race in risk assessment. First, however, we situate the role of gender and risk assessment in empirical and in legal and moral context.

Empirical Context

That women participate in crime, particularly violent crime, at much lower rates than men is a staple in criminology and has been known for as long as official records have been kept. For example, men constituted 91 percent of homicide offenders in thirteenth century England (Given, 1977), and 89 percent of homicide offenders in twenty-first America (Federal Bureau of Investigation, 2015, Table 42). The earliest scientific review of gender differences in violence (Maccoby & Jacklin, 1974, p. 352) concluded that "[t]he sex difference in aggression has been observed in all cultures in which the relevant behavior has been observed. Boys are more aggressive both physically and verbally. . . The sex difference is found as early as social play begins—at age 2 or 2 ½." A subsequent comprehensive review (Sampson & Lauritsen, 1994, p. 19) similarly concluded that "sex is one of the strongest demographic correlates of violent offending . . . . [M]ales are far more likely than females to be arrested for all crimes of violence including homicide, rape, robbery, and assault." Meta-analytic reviews confirm that sex is a relatively strong demographic predictor of general offending (similar to age, race, etc.), despite its modest absolute level of predictive utility (e.g., $r = .10$; Gendreau et al., 1996). Regarding violence, it is hard to gainsay the conclusion of Gottfredson and Hirschi's classic work, *A General Theory of Crime* (1990, p. 145): "gender differences appear to be invariant over time and space."

Importantly, however, the Panel on Research on Criminal Careers of the National Research Council (Blumstein, Cohen, Roth, & Visher, 1986) decomposed the aggregate crime rate into two component parts: "participation" (i.e., "the percentage of the population that commits crimes"), and "frequency" (i.e., "the rate of criminal activity of those who are active [offenders]" (p. 1). The reason for this decomposition is that each of the two components of the crime rate is relevant to different policy considerations. *Participation* is relevant to social programs to prevent people from becoming involved in crime in the first place (e.g., educational and employment initiatives for disadvantaged youth), while *frequency* is "central to the decisions of the criminal justice system," including sanctioning (p. 1). The principal conclusion of the Panel with respect to gender was:

> The large differences between males and females found in aggregate population arrest rates appear to arise predominantly from differences in participation. If active in a crime type [e.g., burglary], females commit that crime at rates similar to those of active males" (Blumstein et al, 1986, p. 67).

Since criminal sanctioning is imposed only on individuals who have already been convicted of participation in crime, the frequency with which an offender continues to commit crime (i.e., the likelihood of recidivism) is the *only* empirical component of the aggregate crime rate relevant to risk assessment. And here the gender difference, while statistically significant, is smaller than the vast gender difference in participation in crime. For example, Durose, Snyder, and Cooper (2015, Table 4) studied the recidivism rates of offenders released from prisons in thirty U. S. states in 2005. Seventy-eight percent of male prisoners were re-arrested for non-traffic offenses within five years after release, compared with 68 percent of female prisoners.

Legal and Moral Context

Leading critics of the use of risk assessment to inform criminal sanctioning have been legal scholars. In the case of race, the focal criticism has been that that risk assessment instruments are biased because—even though they omit race as an explicit risk factor—they include potential correlates of race (e.g., marital history, employment status). In contrast, in the case of gender, the focal issue for legal scholars has been the formal incorporation of gender as a risk factor in assessment instruments used in criminal sanctioning (Desmarais et al, in press). Sonja Starr (2014, p. 806), for example, has written that reliance on gender in sentencing "amounts to overt discrimination" (against men). In her reading, "the Supreme Court's cases on gender… have consistently held that disparate treatment cannot be justified based on statistical generalizations about group tendencies, even if they are empirically supported" (p. 807).

Similarly, Michael Tonry (2014, p. 171) has argued in moral terms against the inclusion of gender and other demographic risk factors in the sanctioning process: …there is something fundamentally unethical or immoral about apportioning punishments or other intrusions on liberty on the basis of ascribed characteristics for which no coherent argument can be made that offenders bear personal responsibility for them. But the legal academy is not of one mind about the use of gender in criminal sanctioning. Richard Frase (2014, p. 149) has noted that the use of gender as a risk factor for recidivism raises different issues than the use of race:

> Since risk-based mitigation of sentences for women… results in more favorable treatment for [a group that has] traditionally been victims of discrimination, there is little reason to fear that differential treatment is motivated by bias or would worsen inequality; nor is the premise for such treatment—lower risk—likely to reinforce negative stereotypes about these groups.

Formal legal policies reflect the stark lack of scholarly consensus on the use of gender as a risk factor for recidivism. While the federal Sentencing Guidelines (U. S. Sentencing Commission, 2014) expressly forbid taking gender into account in determining sentence length, the highly influential Model Penal Code (American Law Institute, 2012) explicitly endorses reliance on gender as a risk factor.

Lay opinion on the use of gender as a risk factor in criminal sentencing appears to cleave as sharply as academic commentary. In a recent survey of American adults, Scurich and Monahan (in press) found that approximately half the respondents were open to the possibility of using gender as a risk factor in sentencing, and half were not.

Bringing Psychological Science to the Controversy

Although the PCRA omits gender as an explicit risk factor, it includes potential correlates of gender (e.g., history of violent offending). Whether the PCRA is subject to gender bias is an empirical question. Ample guidance on test fairness is available from similar efforts undertaken in more mature fields (e.g., cognitive tests used to inform high-stakes education and employment decisions, see Reynolds 2000; Sackett, Borneman & Connelly, 2008). The criteria that indicate when a test is biased have been distilled in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014)—which we refer to as the "Standards."

Given that the raison d'etre for risk assessment instruments is to predict recidivism, the paramount indicator of test bias is *predictive bias* (also known as "differential prediction;" Standard 3.7). On utilitarian grounds alone, any instrument used to inform sentencing must be shown to predict recidivism with similar accuracy across groups. If the instrument is unbiased, a given score will also have the same meaning regardless of group membership (e.g., an average risk score of X will relate to an average recidivism rate of Y for *both* men and women). This is commonly tested by examining whether groups systematically deviate from a common regression line that relates test scores to the criterion (i.e., whether the groups share intercepts and slopes; Cleary, 1968; see also Sackett & Bobko, 2010).

Given a pool of instruments that are free of predictive bias, some instruments will yield greater mean score differences between groups than others (e.g., men, on average, will obtain higher risk scores than women, or vice versa). These instruments are not necessarily biased: "subgroup mean differences do not in and of themselves indicate lack of fairness" (Standard 3.6, p. 65). The notion that mean differences are indicative of test bias is unequivocally rejected in the professional literature because group differences in scores may reflect true differences in recidivism risk, based on group variation "in experience, in opportunity, or in interest in a particular domain" (Sacket et al., 2008, p. 222). Gender is associated with differences in biology, socialization, and sex roles (e.g., Wood & Eagly, 2012) that can relate to risk-taking and criminal behavior. Gender differences in such circumstances can manifest as valid group differences in risk scores.

Even if mean score differences between men and women do not reflect test bias, using instruments that yield such differences to inform sentencing may create *disparate impact* (in legal terms; see *Griggs v. Duke Power,* 1971) or inequitable social consequences (in moral terms; see Reynolds & Suzuki 2012). An instrument can perfectly measure risk, and yet the *use* of the instrument could still be seen as unfair.

In our view, risk assessment instruments used to inform sanctioning (i.e., decisions about imprisonment, release, community supervision, and risk reduction services) must be empirically examined for both predictive bias and disparate impact. That is, risk assessment must be both empirically valid and perceived as morally fair across groups.

This study is among the first to examine such issues with respect to gender. In a recent meta-analysis, Desmarais, Johnson, & Singh (in press) identified 53 published and unpublished studies of 19 risk assessment instruments used in U.S. correctional settings. They found that only one instrument—the Levels of Services Inventory-Revised (LSI-R)—had been examined in relation to gender in multiple samples (k=7). Results indicated similar levels of predictive utility for male and female offenders (r_{pb}=.25 & .23, respectively). Although this indicates that the *strength* of association between risk scores and recidivism is similar across genders, to our knowledge no studies have tested whether the *form* of this association is the same for men and women. That is, no studies have specifically tested predictive fairness.

Present Study

In this study, we use a cohort of male and female federal offenders to empirically examine the relationships among gender, risk assessment, and recidivism. In the federal system, risk assessment is not used to inform sentencing decisions. Instead, the Federal Post Conviction Risk Assessment or "PCRA" (Johnson, Lowenkamp & VanBenschoten, 2011) is used to inform decisions designed to reduce risk—i.e., to identify *whom* to provide with relatively intensive services (i.e., higher-risk offenders) and *what* to target in those services (i.e., variable risk factors). The PCRA was developed by the Administrative Office of the US Courts (Probation and Pretrial Services Office), and is administered post-conviction, upon intake to a term of supervised release or probation. Given that the PCRA is well-validated and includes major risk factors tapped by many other risk assessment instruments, these federal data are well-suited for addressing three aims with broader implications:

1. To what extent is the instrument—and the risk factors it includes—free of *predictive bias?* We hypothesize that there will be little evidence of test bias by gender.
2. To what extent does the instrument yield average score differences between gender groups that are relevant to *disparate impact*? We hypothesize that women will obtain lower PCRA scores than men.

3. Which risk factors contribute the most to mean score differences between men and women? Given past research and the PCRA's scoring system, we expect criminal history to contribute the most to these differences.

METHOD

Participants

Participants were 14,310 offenders drawn from a larger dataset on over 96,000 offenders assessed between August 2010 and November 2013 (see Walters & Lowenkamp, 2015). Because even trivially small differences can become statistically significant in samples as large as ours (Lin, Lucas & Shmueli, 2013), we use an alpha level of .001 to signal statistical significance and focus on effect sizes in interpreting results.

Offenders were selected for inclusion in the current study if they met the following criteria: (a) assessed with the PCRA at least 12 months prior to the collection of follow-up arrest data (to permit tests of predictive bias: *n* lost = 29,680), (b) no missing data on PCRA items (to permit analyses at the risk factor level; *n* lost = 1,007), and (c) race coded as either African American or non-Hispanic white (to permit relevant racial comparisons*;* *n* lost = 17,238). Application of these criteria yielded an eligible pool of 48,475 offenders of which 15% were female and 85% were male.

Female participants were more likely to be white and were, on average, two years younger than male offenders. To equalize the groups on these variables, female offenders were randomly matched to male offenders on race and age (in years), using ccmatch in STATA (Cook, 2015).[1] The matching process identified a male counterpart for each of the 7,155 female offenders yielding a final sample size of 14,310.

As shown in Table 1, offenders' average age was 38.23 with 41% of the sample being classified as African American. For both male and female offenders, the modal conviction offense was for a drug crime. While a similar percentage of offenders in each group were under supervision for a public order related crime, there are considerable differences in the percentages of males and females under supervision for firearms, white collar, property, violent and sex offenses. The minimum follow up time period was 366 days with a maximum of 1,581 days. The mean follow up time period was 1030 days for females and 1038 days for males. There were no significant differences between male and female offenders in their mean follow up period, $t(14292.5) = -2.07$, $p = 0.04$.

Measures of Risk

The history, development, and predictive utility of the Post Conviction Risk Assessment are detailed elsewhere (see Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011; Lowenkamp et al., 2013; Lowenkamp, Holsinger, & Cohen, 2015). The PCRA is an actuarial instrument that explicitly includes variable risk factors and was constructed and validated on large, independent samples of federal offenders. Items that most strongly predicted recidivism in the construction sample contribute most strongly to total scores (Johnson et al., 2011). Fifteen items are scored and weighted (all items are weighted 1 except for the number of misdemeanor and felony arrests (where 0 = none, 1 = one or two, 2 = three through seven, and 3 = eight or more) and age in years at intake to supervision (where 0 = 41 and above, 1 = 26 to 40, and 2 = 25 or younger). Each of the fifteen items is nested under one of five domains—criminal history, employment, social networks, substance abuse, and attitudes. With the exception of criminal history, the PCRA domain scores are changeable over time. The items are summed to yield a total PCRA risk score from 0 to 18. Total PCRA scores place an offender into a risk category: low (0-5), low/moderate (6-9), moderate (10-12), or high (13-18).

The PCRA has been shown to be reliable and valid. Specifically, officers must complete a training and certification process to administer the PCRA. The certification process has been shown to yield high rates of inter-rater agreement in scoring (Lowenkamp et al., 2013). The accuracy of the PCRA in predicting recidivism rivals that of other well-validated instruments (for a review, see Monahan & Skeem, 2014). For example, based on a sample of over 100,000 offenders, Lowenkamp et al. (2015) found that the PCRA moderately-to-strongly predicted both re-arrest for any crime and re-arrest for a violent crime, over up to a two-year period (AUCs=.70-.77). Finally, scores on the PCRA have been shown to change over time. Overall, 18% of the offenders have a change in risk classification from the initial assessment to the first re-assessment. Higher percentages (≈ 40%) of moderate and high-risk offenders are categorized in risk classes that differ from the initial assessment. These changes in classification are associated with changes in the probability of recidivism (Cohen, Lowenkamp, & VanBenschoten, 2015).

The PCRA was administered by officers when an offender entered supervision or when reassessing an offender. In the present study, the results of the earliest assessment were selected for analyses as this provided the longest follow up time period. In addition to the total PCRA score, the sub-scores from the five PCRA domains were also calculated and analyzed.

Arrest Criterion

Data from the National Crime Information Center (NCIC) and Access to Law Enforcement System were used to collect information on arrests. A standard criminal history check was retrieved on each participant that yielded their entire criminal history. The date and types of arrests that occurred after the date of PCRA administration were coded from these data.

The result was two dichotomous measures that we used in analyses of predictive fairness: arrest for any offense (excluding technical violations of standard conditions of supervision), and arrest for any violent offense. Violence was defined using the NCIC definitions (i.e., homicide and related offenses, kidnapping, rape and sexual assault, robbery, assault).

In the present sample, the base rate for any arrest was 24% (18% females; 29% males, $\chi^2(1) = 261.31$; $p \leq 0.001$; $\phi = 0.14$), and the base rate for violent arrest was 5% (3% females; 8% males, $\chi^2(1) =182.75$; $p \leq 0.001$, $\phi = 0.11$). As indicated by the data, males are at considerably greater risk than females of being arrested for any offense and for a violent offense (odds ratio of 1.91 and 2.92 for any and violent re-arrests, respectively).

Analyses

To address our aims, we calculated descriptive statistics, effect sizes (Cohen's *d*), and measures of predictive validity (AUCs and DIF-R; Silver, Smith & Banks, 2000). We also performed regressions to test whether gender moderated the predictive utility of the PCRA.

RESULTS

Testing Predictive Fairness

Our first aim was to test the extent to which the PCRA—and the risk factors it includes—are free of predictive bias by gender. We hypothesized that there would be little evidence that the PCRA's accuracy in predicting arrest differed for men and women. As shown below, results are only partly consistent with this hypothesis: Although the PCRA strongly predicts recidivism for both genders, it tends to overpredict recidivism for women (i.e., predicted probabilities of recidivism associated with PCRA scores—which are based on both genders—overestimate arrest rates for women).

Degree of prediction as a function of gender. We began our analyses by examining whether the *degree* of relationship between PCRA scores and arrest varied as a function of gender (see Arnold, 1982). Table 2 presents arrest rates for offenders classified in each PCRA risk classification (low to high) by gender. Results indicate that rates for any arrest and for violent arrest increase monotonically as risk classifications increase. These findings do not differ across gender subgroups.

[Insert Table 2]

Table 2 also presents DIF-R and AUC values for the overall sample and by gender. The Dispersion Index for Risk (DIFR; see Silver, Smith & Banks, 2000) roughly indicates the extent to which PCRA risk classifications create reasonably sized groups of offenders with arrest rates that are as different as possible. DIFR ranges from 0 to infinity, increasing as the classification model disperses cases into subgroups whose baserates of re-arrest are distant from the total sample baserate and whose subgroup sample sizes are large in proportion to the total sample size. Unlike the DIFR (which focuses on PCRA risk classifications), the Area Under the ROC Curve (AUC) focuses on PCRA Total Scores. The AUC is widely used to assess the accuracy of risk assessment tools, partly because its values are not heavily influenced by differences in base rates of offending (unlike correlations; see Babchishin & Helmus, 2015). Use of the AUC in the present study is critical, given that women's baserate of any arrest (18%) was much lower than that of men (29%, $\chi^2(1) = 261.31$; $p \leq 0.001$; $\phi = 0.14$). Minimum AUCs of .56, .64, and .71 correspond to "small," "medium," and "large" effects, respectively (see Rice & Harris, 1995).

As shown in Table 2, AUC values are consistently large, with no significant difference in predictive utility by gender. This indicates that the PCRA strongly predicts any arrest and violent arrest for both men and women. Similarly, DIFR values are consistently in the high range, compared to other risk assessment tools used in practice (see Skeem et al., 2013)—indicating that the PCRA classifications perform quite well, in terms of base rate dispersion. Although DIF-R rates appear slightly higher for men than women for any arrest,[2] they are very similar across gender for violent arrest.

Form of prediction as a function of gender. Given this finding that the PCRA accounts for roughly the same degree of variance in arrest for men and women, we next examined whether the *form* of the relationship between PCRA scores and arrest varies by gender (see Arnold, 1982). Ideally, an average PCRA score of X will relate to an average arrest rate of Y for both men and women.

To test whether the form of the relationship between the PCRA and arrest was similar by gender, we estimated a series of bivariate logistic regression models (four models for any arrest; four models for violent arrest). These models were compared to test for "subgroup differences in regression slopes or intercepts, [which] signal predictive bias" (SIOP, 2003; see also Aguinis, Culpepper & Pierce, 2010). As shown in Table 3, in Models One and Two, only gender and only the PCRA total score, respectively, were used to predict any arrest. Model Three included both gender and the PCRA, and Model Four included gender, the PCRA, and the interaction between gender and PCRA.

[Insert Table 3]

The results for any arrest are summarized in Table 3—and convey two main findings. First, the slope of the relationship between PCRA scores and any arrest is similar for men and women. Specifically, comparison of Models Three and Four indicate that gender does not significantly moderate the predictive utility of the PCRA, $\Delta\chi^2$ (1) = 1.16, *ns*. Similar results were obtained for violent arrest, $\Delta\chi^2$ (1) = -0.44, *ns*. Second, the intercept of the relationship between

PCRA scores and any arrest is higher for men than for women. That is, comparison of Models Two and Three indicate that gender adds small, but significant incremental utility to the PCRA in predicting both any arrest ($\Delta\chi^2$ [1] = 97.97, *p*< .001) and violent arrest ($\Delta\chi^2$ [1] = 102.18, *p* < .001). The odds ratios from Model 3 indicate that—after taking PCRA total scores into account—men are still 1.53 and 2.27 times more likely than women to be arrested for any crime or a violent crime, respectively. The gender effect is meaningful, for violent arrests.

To concretize gender differences in the form of the relation between the PCRA and arrest, we calculated predicted probabilities of any arrest and for violent arrest based on Model Three (which includes both gender and PCRA scores). We grouped the predicted probabilities together for each PCRA score and plotted the average for each score. As shown in Figure 1, the slope of the relationship between the PCRA and arrest does not differ by gender (although the distance between lines representing men and women may appear to increase with PCRA scores, the distance is actually constant or even decreasing, relative to the base rates of the outcome criterion). At the same time, the intercept of the relationship between the PCRA and arrest is lower for women than men, which translates to overestimation of recidivism for women.

[Insert Figure 1]

Exploring predictive fairness at the domain level. Even in the presence of little predictive bias at the global level for PCRA total scores, individual risk domains may be more- or less- gender neutral. Given findings from the gender specific risk assessment literature (see discussion), we conducted analyses that parallel those described above to test whether the degree and form of the relationship between PCRA domains and arrest differed as a function of gender.

Table 4 presents the degree of association (i.e., AUCs) between PCRA domains and any arrest for the entire sample as well as by gender. Results indicate that there are no significant differences between men and women in the degree of association between each of the risk domains and both any arrest and violent arrest. Criminal history had a large effect and the remaining domains (i.e., employment, substance abuse, social networks, and attitudes) had a small effect.

[Insert Table 4]

Table 5 presents the results of a series of logistic regression models that test the gender fairness of each PCRA domain in predicting any arrest. The four models completed for each domain parallel the models described above for PCRA Total scores. Results are presented in Table 5—with "any arrest" in the top panel and "violent arrest" in the lower panel. For each of the five risk domains, tests of similarity in slopes appear on the left and tests of similarity in intercepts appear on the right.

[Insert Table 5]

With respect to slopes, the left columns present (a) two comparisons of Model 3 (PCRA domain & gender) and Model 4 (PCRA domain, gender, and their interaction), i.e., the change in pseudo-R^2 and change in X^2 and significance level of that change, and (b) for Model 4, the odds ratio of the interaction term. Across domains, model comparisons indicate that gender does not significantly moderate the relation between the risk domain and any- or violent- arrest.—the slopes are similar. Shifting to intercepts, the right columns present (a) two comparisons of Model 2 (PCRA domain only) and Model 3 (PCRA domain & gender), i.e., the change in pseudo-R^2 and change in X^2 and significance level of that change, and (b) for Model 3, the odds ratio and significance level of gender (indicating its incremental utility over PCRA scores). These model comparisons indicate that all of the PCRA domains overestimate rates of any arrest and violent arrest for women, compared to men. The odds ratios indicate that—after taking

PCRA domain scores into account—men are still more likely to be arrested than women. Gender effects are smallest for any arrests and for the domain of criminal history. For example, after controlling for criminal history scores, men are only 1.30 times more likely to be arrested for any offense than women (a trivial effect). But after controlling for social network scores, men are 3.09 times more likely to be arrested for a violent offense than women (a large effect).

Summary. Overall, results are only partly consistent with our expectation of predictive fairness by gender. There is a strong *degree* of association between the PCRA and future arrests for both men and women. But there are some gender differences in the *form* of this association. Although gender does not moderate the relationship between PCRA scores and re-arrest (i.e., slopes are similar), PCRA scores overestimate rates of recidivism for women (i.e., intercepts are lower for women).

Assessing Mean Score Differences Relevant to Disparate Impact

Our second aim was to assess the extent to which the PCRA yields average score differences by gender relevant to *disparate impact*. We hypothesized that women would obtain lower PCRA total scores than men. Results are consistent with that hypothesis.

Specifically, average PCRA total scores for women and men were 5.78 and 6.85, respectively—a raw difference of one point on an 18-point scale that translates into a *d* value of 0.32. According to conventional classifications, minimum *d* values of .2, .5, and .8 define small, medium, and large effects, respectively (Cohen, 1988). A *d* of .30 corresponds to 79% overlap (and 21% non-overlap) between genders in PCRA scores (see Cohen, 1988). The difference is small but possibly meaningful.

Identifying Risk Domains That Underpin Mean Score Differences

Domain differences. Our third aim was to determine which risk factors contribute the most to mean score differences between men and women. We expected criminal history to contribute the most to these differences. Results were consistent with this expectation.

[Insert Table 6]

Mean scores and standard deviations for PCRA risk domains are reported by gender in Table 6, along with Cohen's *d*. Column 8 indicates the percentage of the difference in the PCRA total means that is attributable to a given risk domain. As shown in that column, essentially all of the difference in PCRA scores is attributable to criminal history. Average criminal history scores for female and male offenders were 3.29 and 4.41, respectively—which translates into a medium effect size (*d* = 0.51). Gender differences in scores across the other risk domains were smaller than 'small,' in terms of *d* values. If not for the criminal history domain, PCRA scores for men and women would be nearly identical.

A closer look at the criminal history domain. Criminal history both explained gender differences in PCRA scores *and* was strongly predictive of recidivism regardless of gender. Because criminal history is a ubiquitous risk factor that can be measured in myriad ways are more- or less- related to demographic characteristics (see Frase et al., 2015), we examined the items of the criminal history domain to identify specific differences by gender. In Table 7, we display mean score differences by gender for five of the six criminal history items (age was excluded, given that it was a matching variable).

[Insert Table 7]

As shown in Table 7, differences between men and women across most criminal history items were small (*d* = 0.31-0.40). The one exception was a history of violent offenses, which had a moderate effect (*d*=.54). A history of violent offenses was present for 46% of men and only 22% of women.

Putting Predictive Fairness and Mean Score Differences Together

Figure 2 provides a visual summary of the study's main findings. The bar chart displays arrest rates for any offense by PCRA classification (low to high) and by gender. The line graph in the figure displays the percentage of offenders within each PCRA classification (e.g., "low") that is male. First, the line graph indicates that the percentage of offenders who are male increases modestly, as PCRA risk classifications increase (e.g., 42% of low-risk offenders are male whereas roughly 69% of high-risk offenders are male). This indicates that men tend to obtain somewhat higher PCRA scores than women. Second, the bar chart indicates that (a) rates of arrest increase steeply and systematically by PCRA risk classification for both men and women, and (b) rates of arrest within each PCRA risk classification are consistently lower for women than men (e.g., of offenders classified as moderate risk, 52% of men and 37% of women are re-arrested). This indicates that the PCRA strongly predicts arrest for both men and women, but rates of recidivism for women are consistently lower than those of men. The similar pattern of findings for violent arrest is displayed in Figure 3.

[Insert Figures 2 and 3]

Figure 4 presents a more dimensional and integrated visual summary of the findings. In this figure, we plotted PCRA total scores on the X axis. Percentages of men and women arrested for any crime and for violent crime are plotted on the left Y axis, while the number of men and women with each PCRA total score are plotted against the right Y axis. This Figure displays two of our principal findings: (1) the small (approximately 21%) degree of non-overlap in the distributions of risk scores by gender; and (2) for both any arrests and for violent arrests, men and women's arrest rates by PCRA total scores are similar in shape, but different in elevation. That is, women are generally arrested at a lower rate than men with whom they share a PCRA total score.

DISCUSSION

. Our results make two major points. First, given that risk assessment is relevant only to utilitarian sanctioning goals of crime control, turning a blind eye to the effects of gender on recidivism can translate to discrimination against women. The PCRA—like many other risk assessment instruments and all sentencing guidelines—omits gender. We found that the PCRA over-estimates women's likelihood of recidivism. This is true even through the PCRA strongly predicts recidivism for both women and men—and even though gender does not moderate the relation between the PCRA and recidivism. Given a particular PCRA score, men are 1.53 times more likely than women to be arrested for any crime—and 2.27 times more likely to be arrested for a violent crime. Second, setting aside the issue of test bias, men obtained slightly higher average scores on the PCRA than women—entirely as a function of men's greater criminal history. Given that criminal history is emphasized in most sentencing guidelines, it is not clear that use of the PCRA would increase any disparate impact on men.

Before contextualizing these findings, we note study limitations that should be borne in mind. First, this study is among the first of its kind, so the generalizability of its results to other contexts is unclear. Surprisingly few risk assessment instruments have been tested for predictive bias and mean score differences by gender. Additional research is needed to establish the extent to which our findings generalize from federal- to state-level offenders, and from the PCRA to other risk assessment instruments. The LSI-R—which excludes gender like the PCRA—has been most heavily studied in non-federal correctional contexts. A recent meta-analysis indicated that the predictive utility of the LSI-R is similar for men and women (Desmarais et al., in press). Although this lends confidence to our basic results for the PCRA, it is unclear whether LSI-R (like the PCRA) over-predicts women's recidivism. More importantly, it is unclear whether instruments that *include* gender as a risk factor are less subject to over-prediction than those that exclude gender. Third, data on interrater reliability in scoring the PCRA for this study are not available (but see Lowenkamp et al., 2013 for relevant evidence). Although some risk domains may have been scored more accurately than others, all officers who complete the PCRA must complete a certification process that has been shown to yield reliable scores.

Empirical Issue: Achieving Predictive Fairness for Women

We hypothesized that the PCRA—and the risk factors it includes—would be free of predictive bias. Results are only partly consistent with this hypothesis. The degree of association between the PCRA and recidivism is similar across genders, but the form of that association differs by gender.

Strong degree of predictive utility for women. With respect to the *degree* of association, PCRA total scores strongly predict any arrest and violent arrest for both men and women. Even at the level of specific risk domains (e.g., criminal history, social networks), predictive utility does not differ significantly by gender. Moreover, for both genders, rates of any arrest and violent arrest increase monotonically as risk classifications increase (from low to high). In short, although it was not designed specifically for female offenders, the PCRA "works" for women, in that it strongly predicts their recidivism. These results are consistent with those observed for the LSI-R, a leading instrument developed and validated with predominantly male samples (Smith, Cullen & Latessa, 2009).

Even if risk assessments created for men predict women's recidivism, however, "it is not clear that they would be the assessments we would have, if we had started with women instead of men" (Van Voorhis, Salisbury, Wright, & Bauman, 2010, p. 263). Although risk factors are largely gender neutral, women's pathways into and out of crime may not be identical to those for men (see Kruttschnitt, 2013). There is some evidence that "gender-responsive" risk factors like relationship dysfunction, self-efficacy, and parental stress occasionally add weak incremental utility to the LSI-R, in predicting women's recidivism (Van Voorhis et al., 2010). It is important to recognize that we did not compare gender-neutral and gender responsive assessment approaches.

Over-estimation of recidivism rates for women. Although there is a strong degree of relationship between PCRA scores and recidivism for both women and men, the *form* of that relationship differs by gender. We conducted a series of regression models that indicated that an average PCRA score of X does *not* relate to an average arrest rate of Y for both men and women. The same slope describes the relationship between PCRA total scores and arrest for men and women (i.e., gender does not moderate the relationship), but the intercepts for that relationship are higher for men than for women (i.e., gender adds incremental utility to the PCRA in predicting recidivism). Similar findings are observed across each of the specific PCRA domains.

The fact that men's and women's arrest rates by PCRA scores are similar in shape but different in elevation means that women are generally arrested at a lower rate than men with whom they share a PCRA score and classification. For example, among offenders who share a "moderate" PCRA risk classification, women are arrested for a violent crime at less than half the rate of men (8% and 17%, respectively). This finding lends credence to scholars' and administrators' early impressions that risk assessment instruments tended to "over-classify" women as high risk (see Van Voorhis et al., 2010).

Taking a step back, however, our findings as a whole are inconsistent with legal scholars' categorical arguments that the use of risk assessment instruments to inform sanctioning decisions will have discriminatory effects. We have tested the predictive fairness of the PCRA with respect to two factors that this instrument (like many others) omits—gender and race (Skeem & Lowenkamp, 2015). Unlike race, we found that gender "matters" for predictive fairness. Specifically, the PCRA strongly predicts re-arrest for both Black and White offenders and for both men and women. But a given PCRA score has the same meaning across groups—i.e., same probability of recidivism—only for Black and White offenders. Unless gender-specific recidivism rates are considered when interpreting PCRA scores (see below), the instrument will overestimate women's probability of recidivism.

Our general findings indicate that research can inform legal debate about the use of risk assessment in sanctioning. Demographic factors can relate differently to risk assessment—and risk is directly relevant to the sanctioning goal of preventing new offenses (despite its irrelevance to moral blameworthiness). Lumping together race and gender as risk factors for recidivism—and then avoiding both—can be costly, as observed by Daly & Tonry (1997):

> "In every jurisdiction that changed its sentencing policies [in the 1970s] and attempted to establish sentencing guidelines, [several] propositions were taken as self-evident. First, race and gender were believed to be illegitimate considerations in sentencing… Gender was largely absent from the debates and calculations: if race was a forbidden consideration, so self-evidently was gender. Equal treatment was (and is) a seductive

> criminal justice ideology; there appeared to be no legal or policy alternative… This translated to harsher sentences for women" (pp. 205-206).

Our results suggest that the PCRA's omission of race as a risk factor does not compromise predictive fairness (Skeem and Lowenkamp (2015), but its omission of gender as a risk factor may translate to over-estimation of women's likelihood of recidivism—and potentially harsher sanctions for women.

Moral Issue: Does Risk Assessment Have a Disparate Impact on Men?

Predictive fairness can be tested empirically. Disparate impact cannot. An instrument can perfectly measure risk across groups (i.e., be free of bias)—and yet the *use* of that instrument can be deemed morally unfair because one group obtains higher average scores and is thought to suffer inequitable social consequences. Average score differences between groups are relevant to—but do not establish—disparate impact.

We hypothesized that women would obtain lower PCRA scores than men—and that criminal history would contribute the most to this difference. Results were consistent with both hypotheses. Specifically, average PCRA total scores for women and men were 5.78 and 6.85, respectively—a raw difference of one point on an 18-point scale. This gender difference in PCRA total scores is "small" (*d*= 0.32; appx. 79% overlap between groups). Essentially all of this difference was attributable to criminal history. Average criminal history scores for female and male offenders were 3.29 and 4.41, respectively—which translates into a medium effect size (*d* = 0.51). Gender differences in scores across the other risk domains were smaller than "small" in terms of *d* values. If not for the criminal history domain, total PCRA scores for men and women would have been nearly identical. Criminal history both explained the small gender difference in PCRA scores *and* was strongly predictive of recidivism regardless of gender.

The effect of risk assessment on gender disparities in sanctioning will vary not only as a function of the instrument used, but also as a function of the baseline sanctioning context: Risk assessment, compared to what? Given that criminal history is virtually always considered in sentencing decisions, it is not clear that risk assessment would exacerbate any gender disparities. As noted by Frase et al. (2015), "criminal history scores make up one of the two most significant determinants of the punishment an offender receives [the other being the gravity of the conviction offense] in a sentencing guidelines jurisdiction" and "prior convictions are taken into account by all U.S. sentencing systems" (p. 7). However, because criminal history may be operationalized in myriad ways—and sanctioning decisions extend beyond sentencing—research is needed to identify any conditions under which risk assessment contributes to gender disparities in sanctioning. The effect of a given instrument on such disparities will depend on what practices are being replaced.

Implications and Conclusion

Our results suggest that using a gender-neutral instrument like the PCRA to inform sanctioning decisions risks discriminating against women. The simple way to avoid over-estimating women's likelihood of recidivism is to interpret PCRA scores in a gender-specific manner. That is, to explicitly acknowledge that women who score between 10 and 12 on the PCRA, for example, do not present the same "moderate" risk of recidivism as men who score within the same range on that instrument. Instead, "moderate" risk women have a 37 percent of recidivism while "moderate" risk men have a 52 percent chance of recidivism. As Frase et al. (2015) argue, even if gender is a characteristic that an offender cannot control, it may be fair to give women reduced sanctions reflecting their lower risk—as long as sanctions for men "are proportionate to their current offenses and prior convictions" (p. 102). At a minimum, "policy makers need to wrestle with the fact that practices that are… gender neutral, as well as those that are overtly prejudiced, can produce injustice" (Daly & Tonry, 1997, p. 243).

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *The Standards for Educational and Psychological Testing.* Washington, DC: AERA Publications.

American Law Institute (2014). *Model Penal Code: Sentencing (Tentative Draft No. 3).* Philadelphia: American Law Institute.

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in pre-employment testing. *Journal of Applied Psychology, 95,* 648-680

Arnold, H. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior & Human Performance, 29* 143-174.

Arnold, J., & Arnold, L. (2015). Fixing justice in America. *Politico Magazine.* Available at http://www.politico.com/magazine/story/2015/03/criminal-justice-reform-coalition-for-public-safety-116057.html

Babchishin, K. M., & Helmus, L. (in press). The influence of base rates on correlations: An evaluation of proposed alternative effect sizes with real-world dichotomous data. *Behavior Research Methods.*

Blumstein A, Cohen J, Roth JA, Visher CA. 1986. *Criminal careers and "career criminals."* Washington, DC: National Academy Press.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 115-124.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd ed.* New Jersey: Lawrence Erlbaum.

Cook, D.E. (2015). CCMATCH: Stata module to randomly match cases and controls based on specified criteria. Version 1.3. www.Danielecook.com .

Cullen, F. T., Jonson, C. L,, & Nagin, D. S. (2011). Prisons do not reduce recidivism: the high cost of ignoring science. *Prison Journal, 91,* 48S-65S.

Daly, K., & and Tonry, M. (1997). Gender, race, and sentencing. In M. Tonry (Ed). *Crime and Justice*, Vol. 22. Chicago: University of Chicago Press.

Desmarais, S.L., Johnson, K.L., & Singh, J.P. (in press). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services.*

Durose, M., Snyder H.. & Cooper, A. (2015). *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010.* Washington, D.C.: Bureau of Justice Statistics.

Dvoskin, J. A., Skeem, J. L., Novaco, R. W., Douglas, K. S., eds. (2011). *Using social science to reduce violent offending*. New York: Oxford University Press.

Federal Bureau of Investigation (2015). *Crime in the United States 2014.* Washington, DC: FBI.

Frase, R. S. (2014). Recurring policy issues of guidelines (and non-guidelines) sentencing: risk assessments, criminal history enhancements, and the enforcement of release conditions. *Federal Sentencing Reporter, 26,* 145-157.

Frase, R., Roberts, J. V., Hester, R. & Mitchell , K. L. (2015). *Criminal history enhancements sourcebook.* Minneapolis, MN: Robina Institute of Criminal Law and Criminal Justice.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works!. *Criminology, 34*(4), 575-608.

Given, J. (1977). *Society and Homicide in Thirteenth-Century England*. Stanford: Stanford University Press.

Gottfredson, M., & Hirschi, T. (1990*). A general theory of crime*. Stanford: Stanford University Press.

Griggs vs. Duke Power, 401 U.S. 424 (1971).

Holtfreter, K., & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women. *Journal of Contemporary Criminal Justice, 23,* 363-382.

International Centre for Prison Studies (2013). *World prison brief*. London: International Center for Prison Studies. Retrieved 1/8/16 at http://www.prisonstudies.org/world-prison-brief

Lowenkamp, C. T., Latessa, E. J., Holsinger, A. M. (2006). The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime and Delinquency, 52,*77-93.

Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, *75*, 16-29.

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The Federal Post Conviction Risk Assessment (PCRA): A construction and validation study. *Psychological Services*, *10*, 87-96.

Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA Revisited: Testing the Validity of the Federal Post Conviction Risk Assessment (PCRA). *Psychological Services, 12,* 149-157.

Maccoby, E., & Jacklin, C. (1974). *The psychology of sex differences.* Stanford: Stanford University Press.

Monahan, J. (2014). The inclusion of biological risk factors in violence risk assessments. In I. Singh, W. Sinnott-Armstrong, and J. Savulescu (Eds.). *Bioprediction, biomarkers, and bad behavior: Scientific, legal and ethical implications*. New York: Oxford University Press, pp. 57-76.

Monahan, J., & Skeem, J. L. (2014). Risk redux: the resurgence of risk assessment in criminal sanctioning. *Federal Sentencing Reporter, 26,* 158-166.

Monahan, J., & Skeem, J. L. (in press). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology.*

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. Strickland, & C. R. Reynolds (Eds.). *Handbook of Cross-Cultural Neuropsychology.* New York: Springer, pp. 249-285.

Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment: an empirical review and recommendations. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.). *Handbook of psychology: Assessment Psychology (Second edition)*. New York: Wiley, pp. 82-113.

Rice ME, Harris GT. 2005. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law & Human Behavior, 29*, 615-620.

Rushford, M. (2015). Incarceration helped bring crime down. *New York Times,* October 29. Available at http://www.nytimes.com/roomfordebate/2015/10/29/will-crime-rise-if-more-people-are-kept-out-of-prison

Sabol, W.J., West, H.C., & Cooper, M. (2009). *Prisoners in 2008*. Washington, DC: Bureau of Justice Statistics.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63(4)*, 215-227

Sackett, P. R., & Bobko, P. (2010). Conceptual and technical issues in conducting and interpreting differential prediction analyses. *Industrial and Organizational Psychology*, *3*(2), 213-217.

Sampson, R., & Lauritsen, J. (1994). Violent victimization and offending: Individual-, situational-, and community-level risk factors. In A. Reiss & J. Roth (Eds.). *Understanding and preventing violence: Social influences*. Washington, DC: National Academy Press.

Scurich, N., & Monahan, J. (in press). Evidence-based sentencing: public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior.*

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: A comparison of methods. *Criminal Justice and Behavior*, *27*, 733-764.

Simon J. (2007). Rise of the carceral state. *Social Research, 74,* 471-508.

Society for Industrial and Organizational Psychology (2003). Principles for the Validation and Use of Personnel Selection Procedures, 4th ed. Downloaded 10/10/15 from: http://www.siop.org/_principles/principles.pdf

Skeem, J., Barnoski, R., Latessa, E., Robinson, D., & Tjaden, C. (2013). *Youth risk assessment approaches:* Retrieved 10/10/15 from: http://risk-resilience.berkeley.edu/sites/default/files/wp-content/gallery/publications/BairdRebuttal2013_FINALc1.pdf

Skeem, J. L., & Lowenkamp, C. T. (2015). Risk, race, and recidivism: Predictive bias and disparate impact. *Manuscript under review.* Retrieved 10/10/15 from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339

Skeem, J. L. & Monahan, J. (2011) Current directions in violence risk assessment. *Current Directions in Psychological Science, 20,* 38-42.

Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology and Public Policy, 8,* 183-207.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review,* 66, 803-872.

Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter,* 26,167-176.

Travis, J., Western, B, & Redburn S. (2014). *The growth of incarceration in the United States: Exploring causes and consequences*. Washington, DC: National Academy Press.

U. S. Sentencing Commission (2010). *Sentencing Guidelines*. Washington, DC: U. S. Sentencing Commission.

U. S. Sentencing Commission (2014). *Sentencing Guidelines*. Washington, DC: U. S. Sentencing Commission.

Walters, G. D., & Lowenkamp, C. T. (2015). Predicting recidivism with the Psychological Inventory of Criminal Thinking Styles (PICTS) in community-supervised male and female federal offenders. *Psychological Assessment, online first, available:* http://dx.doi.org/10.1037/pas0000210

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in experimental social psychology*,*46*, 55-123.

Zimring, F. (2012). *The city that became safe: New York's lessons for urban crime and its control*. New York: Oxford University Press.

Table 1: Sample Characteristics

Characteristic	All	Female	Male
Age (SD)	38.23(10.86)	38.23(10.86)	38.23(10.86)
% African American	41.26	41.26	41.26
% Conviction offense^			
Drug	42.45	40.39	44.51
Firearms	11.31	4.19	18.44
White Collar	23.55	33.06	14.03
Public Order	6.60	6.64	6.57
Property	7.04	9.36	4.71
Violence	4.17	2.93	5.42
Sex offense	2.24	0.58	3.90
Average follow-up period in days	1034.20 (233.42)	1030.16 (299.52)	1038.24 (237.21)

^ Categories with less than 5% excluded

Table 2. Predictive Utility of PCRA Risk Classifications and Total Scores by Gender

Feature	Any Arrest			Violent Arrest		
	All	Female	Male	All	Female	Male
% Arrested by PCRA Classification						
Low	9.87	8.36	11.91	1.33	0.86	1.97
Low/Moderate	26.45	22.30	30.26	6.00	3.63	8.17
Moderate	46.41	36.96	52.24	13.14	7.66	16.53
High	60.68	52.75	64.27	18.12	12.09	20.84
DIF-R, PCRA Categories	0.85	0.75	0.86	1.05	0.97	0.98
AUC, PCRA Total[1]	0.74	0.72	0.74	0.75	0.74	0.73

[1] Difference is not significant for Rearrest (Z = -1.52; p = 0.13), nor for Violent Rearrest (Z =0.40; p = 0.69)

Table 3. Logistic Regression Models Testing Gender Fairness of PCRA Total Scores in Predicting Any Arrest

	Model 1	Model 2	Model 3	Model 4
Male	1.91*	--	1.53*	1.37
PCRA Total	--	1.32*	1.31*	1.29*
Male X PCRA Total Interaction	--	--	--	1.02
Constant	0.22*	0.04*	0.04*	0.04*

Note: Values are odds ratios for each predictor.
Model 1 Log Likelihood = -7672.15; $X^2(2)$ = 263.40; $p < 0.001$; pseudo R^2 = 0.02; n = 14,310
Model 2 Log Likelihood = -6846.96; $X^2(2)$ =1913.78; $p < 0.001$; pseudo R^2 = 0.12; n = 14,310
Model 3 Log Likelihood = -6798.13; $X^2(2)$ =2011.44; $p < 0.001$; pseudo R^2 = 0.13; n = 14,310
Model 4 Log Likelihood = -6797.55; $X^2(4)$ =2012.60; $p < 0.001$; pseudo R^2 = 0.13; n = 14,310
* $p \leq 0.001$

Table 4. Utility of PCRA Domain Scores in Predicting Any Arrest

	Any Arrest, AUCs[t]			Violent Arrest, AUCs		
	All	Women	Men	All	Women	Men
Criminal History	0.73	0.71	0.72	0.75	0.73	0.73
Employment	0.61	0.60	0.63	0.62	0.64	0.62
Drugs/Alcohol	0.59	0.58	0.59	0.56	0.55	0.57
Social Networks	0.60	0.60	0.60	0.58	0.60	0.59
Attitude	0.55	0.54	0.55	0.55	0.56	0.54

[t] Z comparisons by gender indicate no statistically significant differences at p<.001

Table 5. Logistic Regression Models Testing Gender Fairness of PCRA Domains in Predicting Arrest

	Slope Comparisons (Models 3 vs. 4)			Intercept Comparisons (Models 2 vs. 3)		
	R^2 Change	X^2 Change	OR, Interaction (Model 4)	R^2 Change	X^2 Change	OR, Gender (Model 3)
Any Arrest						
Criminal History	0.00	2.03	1.03	0.00	36.53*	1.30*
Employment	0.01	8.01	1.12	0.03	284.74*	1.99*
Drugs/Alcohol	0.00	0.22	0.97	0.02	234.00*	1.86*
Social Networks	0.00	0.01	1.01	0.03	314.98*	2.06*
Attitudes	0.00	0.93	1.12	0.03	247.96*	1.89*
Violent Arrest						
Criminal History	0.00	0.507	1.03	0.01	55.33*	1.86*
Employment	0.00	0.194	0.97	0.04	197.57*	3.01*
Drugs/Alcohol	-0.01	0.395	1.09	0.04	177.98*	2.84*
Social Networks	0.00	0.795	0.92	0.04	207.51*	3.09*
Attitudes	0.01	2.505	0.72	0.04	181.69*	2.87*

Note: OR=Odds Ratio, with terms representing the unique effect for men compared to women (male=1; female=0)
*p < .001

Table 6. PCRA Total and Domain Scores by Gender

	Women			Men					d		
Variable	N	Mean	Std. Dev.	N	Mean	Std. Dev.	Difference	% Attributable To Each Domain	Estimate	Lower	Upper
PCRA Total	7,155	5.78	3.19	7,155	6.85	3.45	1.06		0.32	0.29	0.35
Criminal History	7,155	3.29	2.14	7,155	4.41	2.30	1.12	1.06	0.51	0.47	0.54
Employment/Education	7,155	1.01	0.99	7,155	0.98	0.98	-0.03	-0.02	-0.03	-0.06	0.00
Drugs/Alcohol	7,155	0.19	0.47	7,155	0.25	0.52	0.06	0.03	0.11	0.08	0.14
Social Networks	7,155	1.19	0.81	7,155	1.08	0.79	-0.11	-0.06	-0.13	-0.17	-0.10
Attitudes	7,155	0.10	0.30	7,155	0.12	0.33	0.02	0.01	0.08	0.05	0.11

Table 7. PCRA Criminal History Item Scores by Gender

	Women			Men				d		
Variable	N	Mean	Std. Dev.	N	Mean	Std. Dev.	Difference	Estimate	Lower	Upper
Prior Arrests	7,155	1.38	1.09	7,155	1.81	1.08	0.44	0.40	0.37	0.44
Violent Offenses	7,155	0.22	0.41	7,155	0.46	0.5	0.25	0.54	0.51	0.58
Varied Offending Pattern	7,155	0.54	0.50	7,155	0.71	0.45	0.17	0.36	0.33	0.39
Community Supervision	7,155	0.3	0.46	7,155	0.45	0.5	0.15	0.31	0.28	0.34
Institutional Misconduct	7,155	0.11	0.32	7,155	0.23	0.42	0.12	0.32	0.29	0.35

Figure 1. Predicted Probabilities of Any Re-Arrest by PCRA Score and Gender

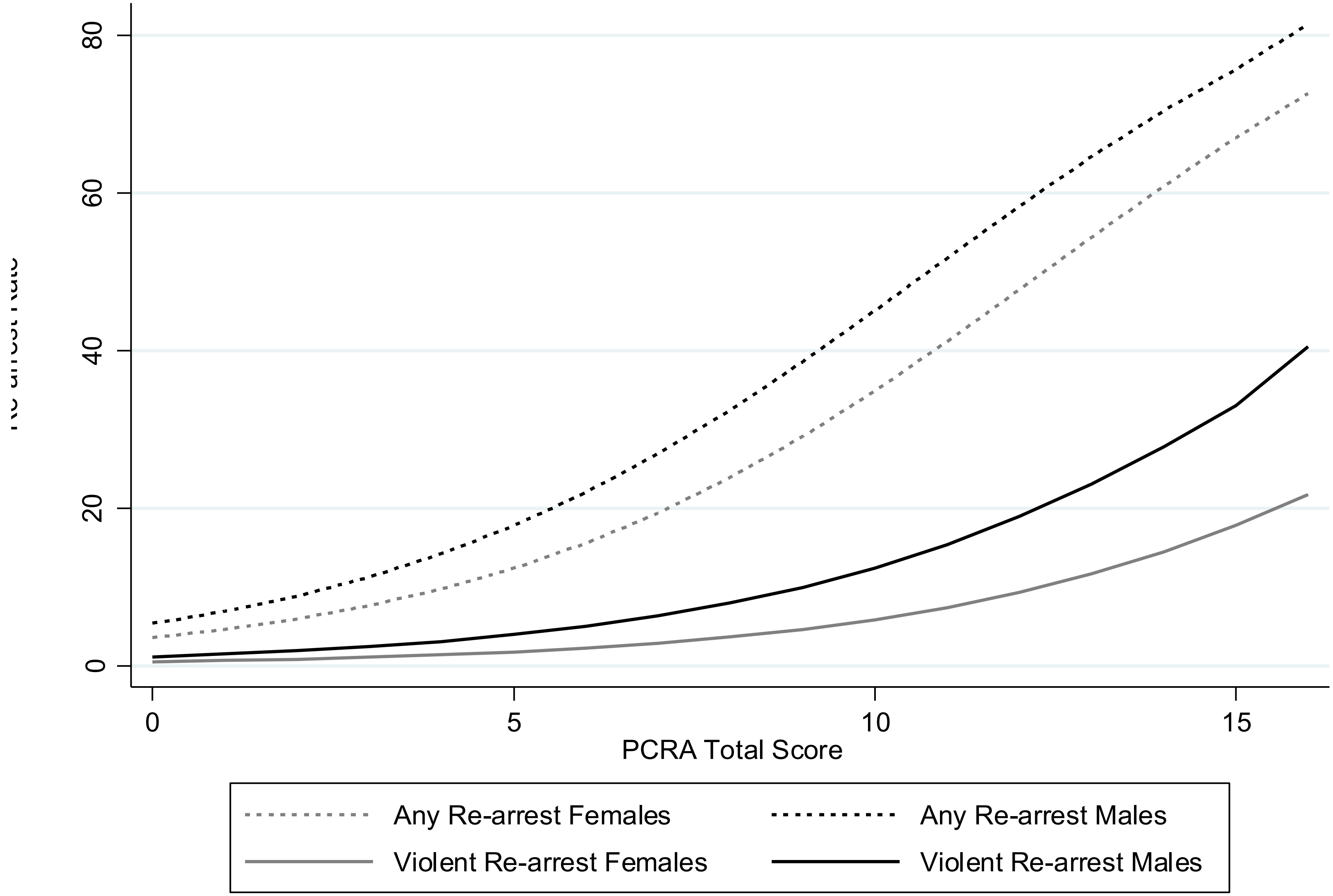


Figure 2. Rate of Re-Arrest for Any Crime and Percent Male by PCRA Score

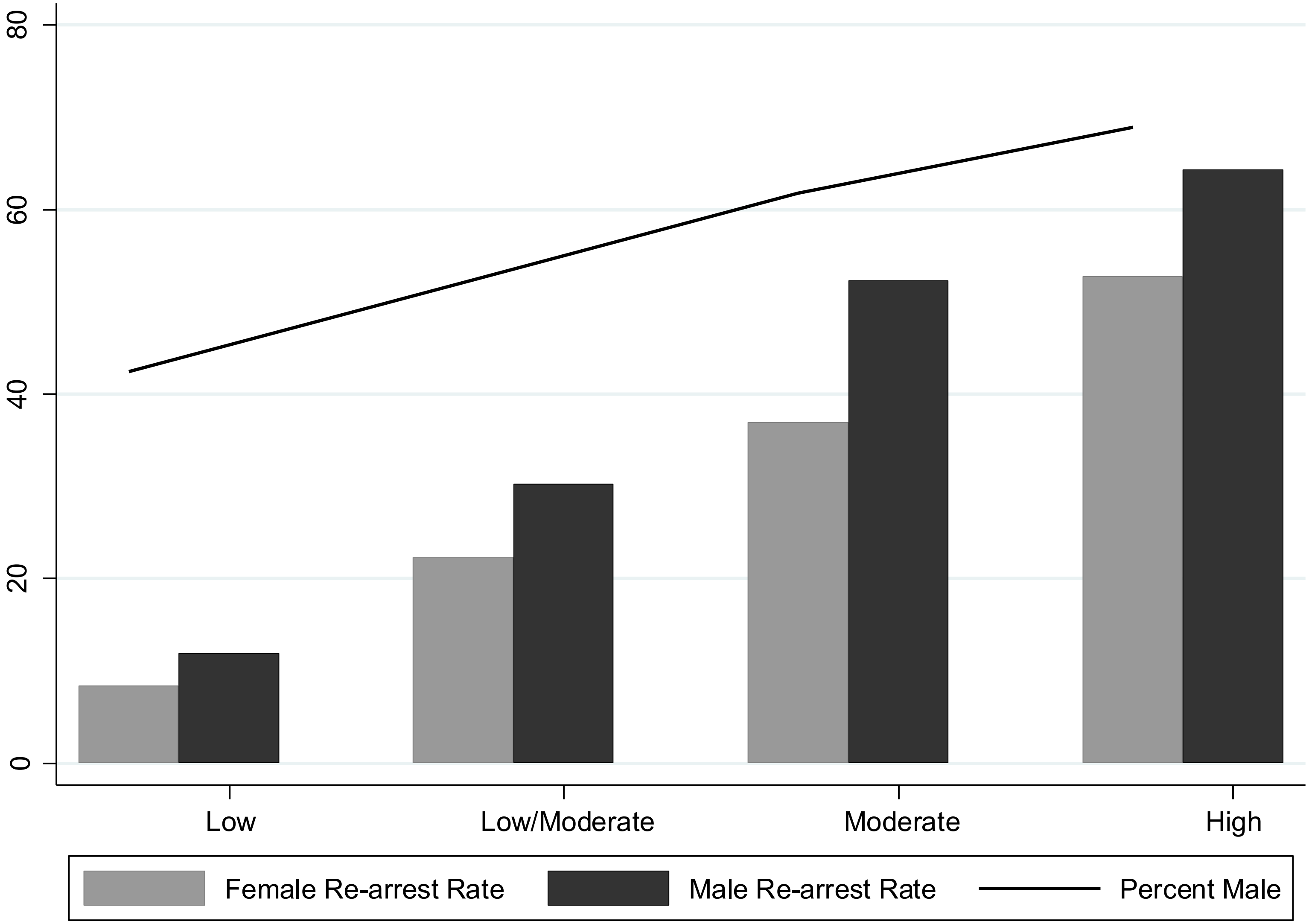


Figure 3. Rate of Re-Arrest for Violent Crime and Percent Male by PCRA Score

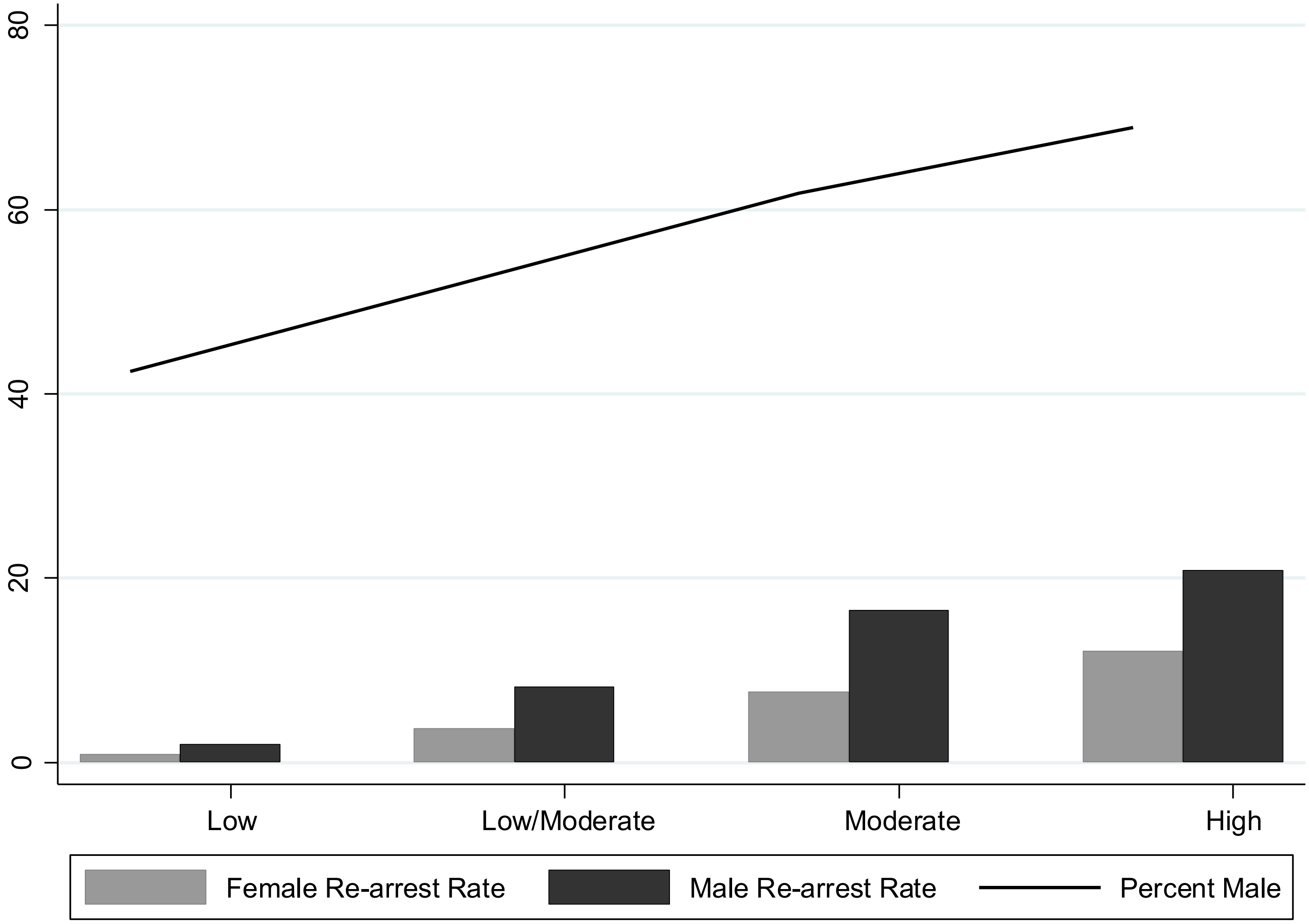
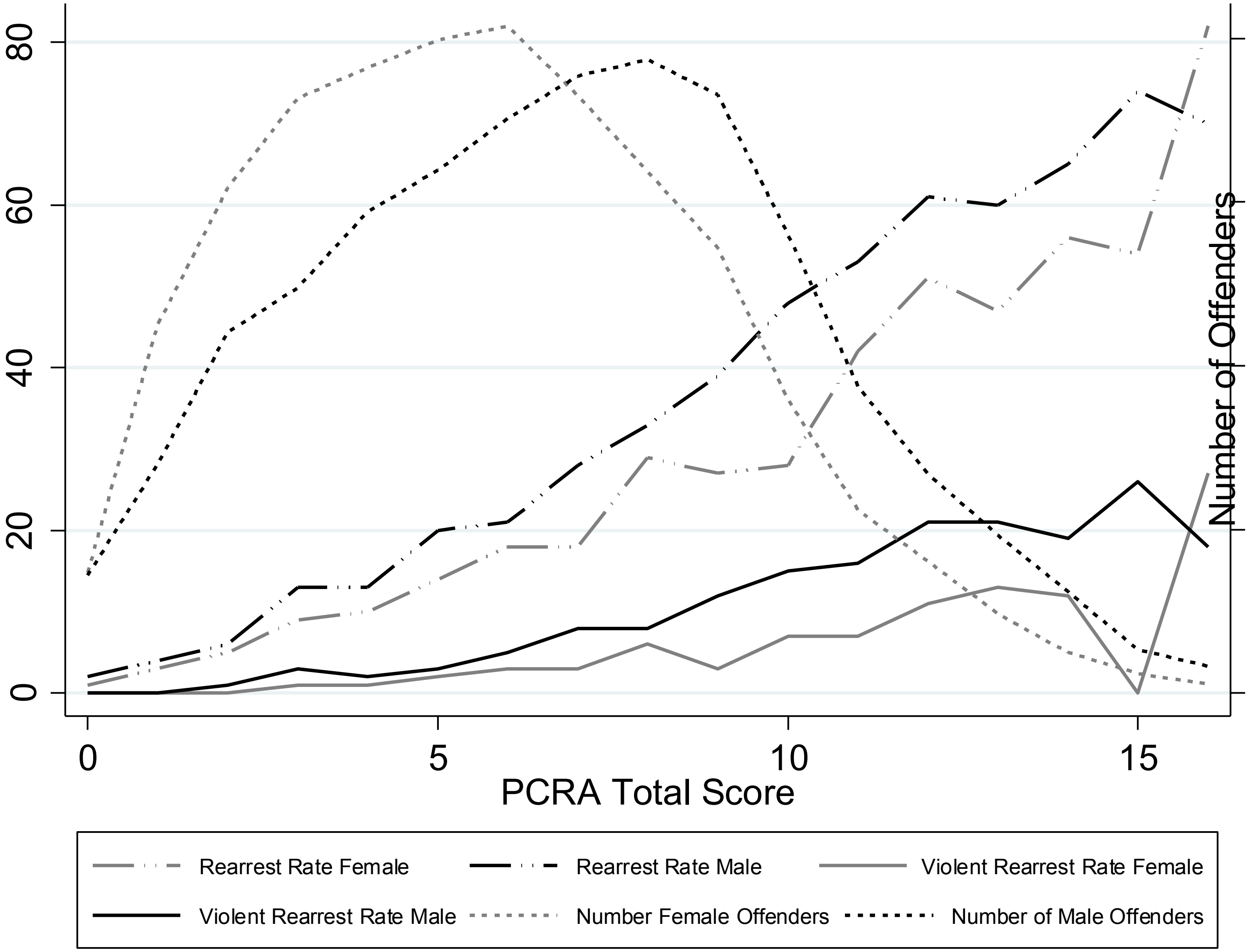


Figure 4. Rate of Re-Arrest for Any- and Violent- Crime and PCRA Distribution by Gender



ENDNOTES

[1] Prior to matching, the average age of female and male offenders was 38.23 and 40.30 respectively ($t(9945.23) = -14.85$; $p \leq 0.001$). Further, 59% of the females were white while 47% of the male offenders were white ($\chi^2(1) = 344.49$; $p < 0.001$). By matching on age and race, we focus more specifically on the relationship between risk and gender.

[2] Because no cutoff values for small, medium, and large values of the DIF-R are available it is not possible to compare them using these benchmarks. Further, since no formulae are available to estimate the confidence intervals of the DIF-R it is not possible to determine if the DIF-R values for male and female offenders differ significantly from one another. Finally, DIF-R values are affected by base rates of recidivism, which vary by gender in this study.